UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Instituto de Estadística

Elaboración De Patrones Espirométricos Normales en Niños Uruguayos Mediante Modelos GAM Y GAMLSS: Parte 1-Identificación de la distribución de la variable de respuesta

Alvarez-Vaz, Ramón
Palamarchuk,Pablo
Riaño, Eugenia

Noviembre,2016

Documentos de Trabajo

Serie DT (16 / 3) - ISSN : 1688-6453

Forma de citación sugerida para este documento:

Alvarez Vaz Ramón, Palamarchuk Pablo,Riaño,Eugenia (2016) 'Elaboración De Patrones Espirométricos Normales en Niños Uruguayos Mediante Modelos Gam y Gamlss:Parte 1-Identificación de la distribución de la variable de respuesta'.
Serie Documentos de Trabajo, DT 3/2016. Instituto de Estadística, Facultad de Ciencias Económicas y Administración, Universidad de la República, Uruguay.

Elaboración De Patrones Espirométricos Normales en Niños Uruguayos Mediante Modelos GAM Y GAMLSS: Parte 1- Identificación de la distribución de la variable de respuesta

Alvarez-Vaz, Ramón [1]

Palamarchuk,Pablo [2]

Riaño,Eugenia [3]

RESUMEN

En un estudio sobre valores de espirometría es necesario identificar un modelo que permita caracterizar curvas percentilares de respuesta espirométricas por edad, sexo y demás características individuales de los participantes. El presente estudio está siendo llevado adelante por un grupo de investigadores del Centro Hospitalario Pereira Rossell, teniendo como población niños de 6 a 12 años, de escuelas públicas y privadas de Montevideo y del interior del país.

En este documento se han utilizado los datos basados en una muestra de aproximadamente 450 niños que, al ser incompleta, no permite determinar la tasa de no respuesta y eventuales sesgos de selección. Aquí se presentan los resultados preliminares, obtenidos mediante métodos de remuestreo, acerca de la identificación de varias familias de distribuciones paramétricas como posibles alternativas para la modelización de las las variables de respuesta. El principal objetivo del estudio es identificar los modelos GAM (General Additive Models) y GAMLSS (Generalized Additive Models for Localization, Scale and Shape), que son un conjunto de modelos de regresión semi-paramétricos que permiten trabajar con una gran cantidad de distribuciones para las variables de respuesta, de tipo discreto, continuo y mixto, con la ventaja de poder considerar distribuciones que presentan censura o truncamiento. Esta clase de modelos se usa en datos de tipo longitudinal, particularmente en las curvas de crecimiento en humanos.

Las técnicas empleadas y su implementación mediante el software R, serán ejemplificadas a través del análisis de la variable de respuesta CVF.

[1] *Instituto de Estadística - Facultad de Ciencias Económicas y de Administración - UdelaR.*

[2] *Estudiante Licenciatura en Estadística- Facultad de Ciencias Económicas y de Administración - UdelaR.*

[3] *Instituto de Estadística - Facultad de Ciencias Económicas y de Administración - UdelaR.*

Palabras clave:Ajuste de distribuciones, Espirometría, Modelos GAM, Modelos GAMLSS, Remuestreo

Códigos JEL:C13,C14,C15,C18
Clasificación MSC2010:62G07,62F40,62J12,62P10

1. Introducción

En un estudio sobre valores de espirometría es necesario identificar un modelo que permita caracterizar curvas percentilares de respuesta espirométricas por edad, sexo y demás características individuales de los participantes.

Fue Hutchinson (Spriggs, 1978) (figura 1) quien desarrolló en el año 1852 el primer espirómetro y las bases de los actuales conceptos sobre la función respiratoria. Desde entonces se han desarrollado sobre estas bases los actuales espirómetros que constan de todos los parámetros clínicos necesarios para interpretar los estudios. Cuando se mide la función pulmonar se identifican parámetros clínicos en el volumen y en el flujo de las respiraciones, tanto en inspiración como espiración.

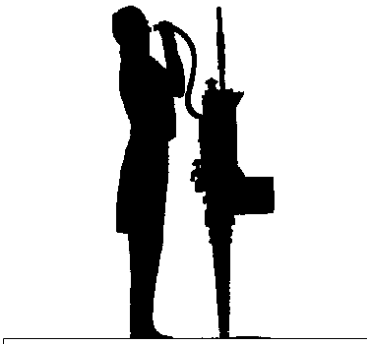

Figura 1: Hutchinson y su espirómetro señalando la posición correcta para la realización de la espirometría.

La maniobra mas relevante es la espiratoria y en forma forzada, partiendo desde una inspiración profunda. Las dos curvas que se presentan en este estudio son: la curva flujo/volumen y la curva volumen/tiempo. El esfuerzo espiratorio máximo nos mide el Volumen Espiratorio Forzado o Capacidad Espiratoria Forzada (CVF), los flujos forzados en 1 segundo (FEV_1) y los flujos forzados denominados periféricos (FEF_{25}, FEV_{50}, FEV_{75} y FEF_{25-75}), que corresponden a porciones de la curva FLUJO/VOLUMEN y representan los flujos de la vía aérea más pequeña.

Además se mide el Indice de Gaënsler - el cual se determina con la espiración forzada expresada a través de un cociente FEV_1/CVF (que se interpreta como % de CVF). El FEV y el CVF deberían ser iguales al realizar el mismo esfuerzo en forma forzada aunque en algunos casos el Gaënsler es menor debido al colapso de la vía aérea durante el esfuerzo y esto sucede en los niños menores (figura 2) .

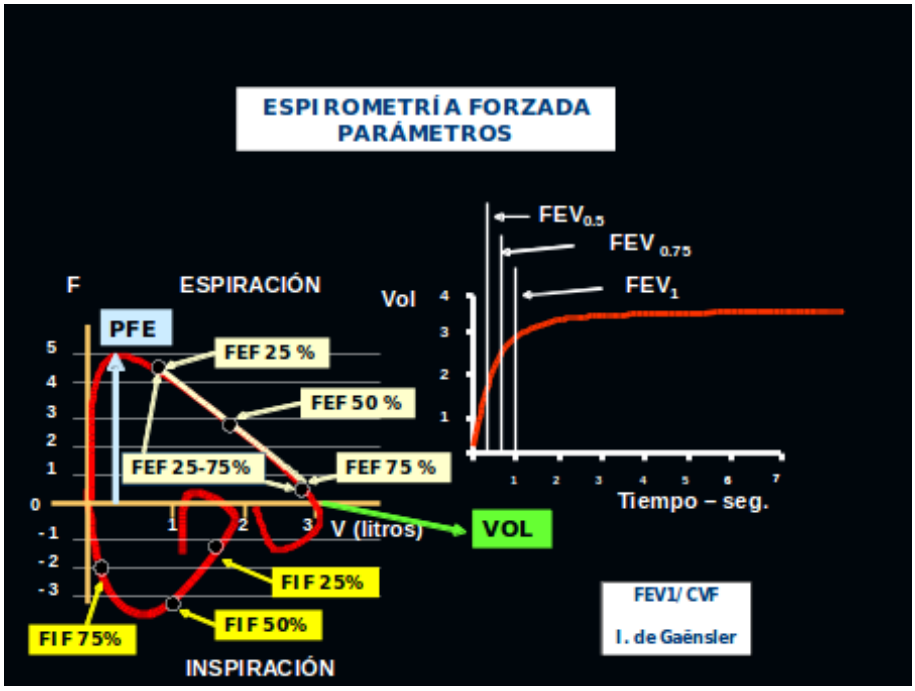

Figura 2: Espirometría Forzada -Parámetros de Curva flujo/volumen y volumen/tiempo

La curva flujo/volumen comienza en el tiempo inspiratorio con una inspiración forzada y continúa con una espiración forzada en el menor tiempo posible. En el tiempo espiratorio tiene un ascenso espiratorio rápido para luego descender en forma progresiva pero más lenta. En la primera parte del ascenso hasta llegar al pico de flujo espiratorio se utilizan todos los músculos espiratorios en esta maniobra por lo que los parámetros que se miden en este tramo son esfuerzo dependientes. Luego de esta primera fase rápida comienza un descenso lento que sí corresponde a los flujos que no dependen de las fuerzas elásticas del pulmón por lo que adquieren importancia en la interpretación de las obstrucciones y restricciones.

La curva volumen/tiempo relaciona el volumen espirado con el tiempo empleado en la espiración. Tiene un ascenso rápido y luego una meseta que se prolonga hasta el final de la espiración. En ella podemos medir Flujos, Volúmenes y el tiempo espiratorio (figura 2).

A continuación se listan los parámetros que mide un espirómetro.

PFE Pico de Flujo Espiratorio. Es el máximo volumen alcanzado en una espiración forzada. Se expresa en L/s (espirómetros) o L/min (medidores portátiles).

FEF Flujo Espiratorio Forzado. Al 25, 50 y 75 % del volumen total espirado, y la porción 25-75 de la misma. Es decir, el flujo máximo cuando resta el 75, 50 y 25 % del volumen a espirar. Se expresa en L/s.

FIF Flujo Inspiratorio Forzado. Al 25, 50 y 75 % del volumen total inspirado. Se expresa en L/s.

Cuando existe Obstrucción de la vía aérea se presenta una disminución de los flujos, tanto del FEV_1 como de los periféricos, manteniéndose la CVF. Cuando existe un atrapamiento de aire o restricción el FEV_1 y el CVF disminuyen proporcionalmente y también la relación CVF/FEV_1, considerándose que existe una Restricción.

La espirometría varía de acuerdo al tamaño de los pulmones. Por lo tanto, los valores varían de acuerdo a la edad, la talla y el peso. Pero también varían de acuerdo a la raza y los diferentes países.

Por esta razón es necesario poseer valores estimados normales para las distintas poblaciones a fin de que los que se apartan de los rangos considerados como normales, puedan ser derivados para su estudio y controlar los tratamientos realizados en ellos. Es por eso que se desarrolla un estudio para medir la función pulmonar en una población de niños uruguayos considerados normales.

En la sección 2 se da cuenta de la metodología estadística usada, en la sección 3 se explica como se llevó adelante el estudio y la generación de los datos; en la sección 4 se presentan los resultados encontrados hasta el momento, los que se discuten brevemente en la sección 5, para terminar en la sección 6, con la indicación de los pasos a seguir.

2. Metodología

El análisis de regresión es una de las técnicas estádisticas más populares y poderosas para la exploración de las relaciones entre una variable de respuesta y sus variables explicativas de interés. Al igual que todos los modelos, los modelos de regresión se basan en ciertos supuestos que necesitan cumplirse (o aproximarse al cumplimiento de los mismos) para que éste tenga conclusiones válidas. Aquellos ususarios que utilizan los modelos de regresión lineal estándar, pronto encuentran que los supuestos clásicos sobre la normalidad y la varianza constante de los errores, y la linealidad de la relación entre la variable de

respuesta y las explicativas, raramente se sostienen.

El modelo lineal ordinario $Y = X\beta + e, Y \sim N(\mu, \sigma^2 I_n)$ donde $\mu = \mathbf{X}\beta$, ha sido extendido por Nelder y Weddeburn (Nelder and Wedderburn, 1972) quienes introdujeron los llamados modelos lineales generalizados (GLM) y los Modelos Aditivos Generalizados (Generalized Additive Models, GAM), que fueron introducidos por Hastie y Tibshirani (Hastie and Tibshirani, 1986) respectivamente para superar algunas de las limitaciones de los modelos lineales estándar.

Desafortunadamente, especialmente con tablas de datos muy grandes, estos modelos se han encontrado con tener ajustes inadecuados o ser inapropiados en gran parte de las situaciones prácticas.

Los Modelos Aditivos Generalizados de Localización, Escala y Forma (Generalized Additive Models for Location Scale and Shape, GAMLSS), son un marco de referencia que corrige algunos de los problemas de los GLM y GAM. Un GAMLSS es un modelo de regresión general, que asume que la variable de respuesta (dependiente), tiene alguna distribución paramétrica. Además, todos los parámetros de la distribución de la variable de respuesta pueden ser modelizados como funciones de las variables explicativas disponibles. Esto contrasta con los GLM y GAM, donde la distribución de la variable de respuesta está restringida a distribuciones de la familia exponencial y sólo se modeliza la media (parámetro de localización) de la distribución.

Entonces, la principal característica de los modelos GAMLSS es la capacidad de permitir que la localización, la escala y la forma de la distribución de la variable de respuesta, varíen de acuerdo a los valores de las variables explicativas.

Los GAMLSS fueron introducidos por Rigby y Stasinopoulos (Rigby and Stasinopoulos, 2005), Stasinopoulos y Rigby (Rigby and Stasinopoulos, 2006) como una forma de superar algunas de las limitaciones asociadas con los modelos lineales generalizados (GLM) y los modelos aditivos generalizados (GAM).

2.1. Modelos Lineales Generalizados

La ecuación para el modelo lineal $\mathbf{Y} \sim N(\mu, \sigma^2 I_n)$, donde $\mu = \mathbf{X}\beta$, permite la generalización para los modelos lineales generalizados (GLM), de acuerdo a Nelder y Weddeburn (Nelder and Wedderburn, 1972). Primero, la distribución normal para la variable de respuesta $\mathbf{Y}_i$, es reemplazada por una distribución de la familia exponencial (EF), y segundo, una función monótona de *enlace* (*link*), $g(\cdot)$ relacionando la media de la variable $\mathbf{Y_i}$, se

considera una función μ_i con el predictor lineal introducido $\eta_i = \mathbf{x_i}^\mathsf{T}\beta$:

$$\mathbf{Y_i} \sim EF(\mu_i, \phi)$$

$$g(\mu_i) = \eta_i = \mathbf{x_i}^\mathsf{T}\beta, \tag{1}$$

independientemente para $i = 1, 2, ..., n$. En notación vectorial un GLM es representado de la forma:

$$\mathbf{Y} \sim EF(\mu, \phi),$$

$$g(\mu) = \eta = \mathbf{x}^\mathsf{T}\beta. \tag{2}$$

La distribución de la familia exponencial $EF(\mu, \phi)$ es definida por la función de probabilidad (densidad) $f_\mathbf{Y}(\mathbf{y}; \mu, \phi)$ de $\mathbf{Y}$ de la forma:

$$f_\mathbf{Y} = exp\left\{\frac{\mathbf{y}\theta - b(\theta)}{\phi} + c(\mathbf{y}, \phi)\right\}, \tag{3}$$

donde $E(\mathbf{Y}) = \mu = b'(\theta)$ y $Var(\mathbf{Y}) = \phi Var(\mu)$ donde la *función de varianza* $V(\mu) = b''[\theta(\mu)]$. La forma de (3) incluye a varias distribuciones importantes, entre ellas la normal, Poisson, gamma, Gaussiana inversa, y también distribuciones binomial y binomial negativa.

2.2. Modelos Aditivos Generalizados

Un Modelo Aditivo Generalizado (Hastie and Tibshirani, 1986) es un modelo lineal generalizado con un predictor lineal que involucra una suma de funciones de suavizado de covariaciones. En general, el modelo tiene la siguiente estructura :

$$g(\mu_i) = \mathbf{X}_i^*\theta + f_1(\mathbf{x}_{1i}) + f_2(\mathbf{x}_{2i}) + f_3(\mathbf{x}_{3i}, \mathbf{x}_{4i}) + ... \tag{4}$$

donde $\mu_i \equiv E(\mathbf{Y_i})$ y $\mathbf{Y_i} \sim$ alguna distribución de la familia exponencial. $\mathbf{Y_i}$ es una variable de respuesta, $\mathbf{X_i^*}$ es una fila de la matriz de modelo para cualquier componente estrictamente paramétrico del modelo, θ es el vector parámetro correspondiente, y la f_j son funciones de suavizado de las covariables $\mathbf{x_k}$. El modelo permite una especificación más flexible de la dependencia de la respuesta en las covariables, pero especificando el modelo sólo en términos de "funciones de suavizado" en lugar de relaciones paramétricas detalladas. Esta flexibilidad y conveniencia conllevan dos nuevos problemas teóricos: la elección de la función de suavizado y el ajuste de sus parámetros (suavidad).

2.3. Modelos Mixtos y Modelos Aditivos Generalizados Mixtos

Un abordaje diferente para estimar e inferir con los GAMs se basa en representar los GAMs como modelos mixtos con los términos de suavizado como *efectos aleatorios* (random effects). Para facilitar la explicación de este abordaje, primero se introducen los modelos mixtos lineales (Wood, 2006), empezando por el modelo mixto simple para datos experimentales balanceados, y luego ir hacia los modelos mixtos lineales generales. Pinheiro y Bates (Pinheiro and Bates, 2000) abarcan la totalidad de la modelización de modelos mixtos lineales en R, mientras que (Ruppert et al., 2003) incluye una explicación clara de las funciones de suavizado como componentes de los modelos mixtos.

En general, los modelos mixtos lineales extienden el modelo lineal ordinario al modelo

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon,\ \mathbf{b} \sim N(\mathbf{0}, \psi), \epsilon \sim N(\mathbf{0}, \mathbf{\Lambda}\sigma^2)$$

donde el vector aleatorio, $\mathbf{b}$, contiene los efectos aleatorios, con media 0 y matriz de varianzas y covarianzas ψ, y $\mathbf{Z}$ es una matriz modelo para los efectos aleatorios. $\mathbf{\Lambda}$ es una matriz definida positiva, de estructura simple, que es típicamente usada para modelar la autocorrelación de los residuos: sus elementos son usualmente determinados por algún modelo simple, con pocos (o sin) parámetros desconocidos. A menudo $\mathbf{\Lambda}$ es simplemente la matriz identidad. Esta extensión permite al modelo tener una estructura estocástica más compleja que el modelo lineal ordinario, y, en particular, implica que los elementos del vector de respuesta, $\mathbf{y}$, ya no son independintes.

2.4. Modelos Aditivos Generalizados de Localización, Escala y Forma

Los Modelos Aditivos Generalizados de Localización, Escala y Forma (GAMLSS) son un tipo de modelos de regresión semi-paramétricos. Son paramétricos, en el sentido que requieren de una suposición de que la variable de respuesta tenga una distribución paramétrica, y "semi" en el sentido de que el modelado de los parámetros de la distribución, como función de las variables explicativas, puede involucrar el uso de funciones de suavizado - *smoothing*- no paramétricas.

En los GAMLSS el supuesto de la familia exponencial para la distribución de la variable de respuesta (Y) es relajado y reemplazado por una distribución general, incluyendo distribuciones continuas y discretas con alto grado de asimetría y/o curtósis. La parte sistemática del modelo es expandida para permitir el modelado, no sólo de la media (o localización), sino que también de otros parámetros de la distribución de Y como función

lineal y/o no lineal, paramétrica y/o suavizados no-paramétricos de las variables explicativas y/o efectos aleatorios. Por lo tanto los GAMLSS están especialmente indicados para modelar una variable de respuesta que no sigue una distribución de la familia exponencial, o que presenta heterogeneidad (por ejemplo, cuando la escala y la forma de la distribución de la variable de respuesta cambian según las variables explicativas).

2.5. El modelo GAMLSS

Un modelo GAMLSS asume que, para $i = 1, 2, ..., n$, observaciones independientes de la variable de respuesta Y_i, tienen función de densidad $f_Y(y_i|\theta^i)$ condicional en $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$, un vector de cuatro parámetros de distribución, cada uno de los cuales puede ser una función de las variables explicativas.

Esto es denotado por $Y_i|\theta^i \sim D(\theta^i)$, o lo que es igual, $Y_i|(\mu_i, \sigma_i, \nu_i, \tau_i) \sim D(\mu_i, \sigma_i, \eta_i, \tau_i)$ independientemente para $i = 1, 2, ..., n$, donde D representa la distribución de Y. Nos vamos a referir a $(\mu_i, \sigma_i, \nu_i, \tau_i)$ como los *parámetros de distribución*. Los primeros dos parámetros de distribución de la población, μ_i y σ_i, se caracterizan normalmente por ser el parámetro de localización y el parámetro de escala, mientras que los restantes, si los hay, son caracterizados como parámetros de forma (asimetría y curtosis).

Sea $\mathbf{Y}^\mathsf{T} = (Y_1, Y_2, ..., Y_n)$ el vector de largo n de la variable de respuesta. Rigby y Stasinopoulos (Rigby and Stasinopoulos, 2010) definen la formulación original de un modelo GAMLSS de la siguiente manera. Para $k = 1, 2, 3, 4$, sea $g_k(.)$ una función de enlace monótona conocida que relaciona el parámetro de distribución $\boldsymbol{\theta}_k$ al predictor η_k:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$$

llevado al caso

$$\begin{aligned}
g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \textstyle\sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1} \\
g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \textstyle\sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2} \\
g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \textstyle\sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3} \\
g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \textstyle\sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4}
\end{aligned}$$

donde $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$, y, para $k = 1, 2, 3, 4$, $\boldsymbol{\theta}_k$ y $\boldsymbol{\eta_k}$ son vectores de largo n, $\boldsymbol{\beta}_k^\mathsf{T} = (\beta_{1k}, \beta_{2k}, ..., \beta_{J'_k k})$ es un vector de parámetros de largo J'_k, $\mathbf{X}_k$ es una matriz de diseño fija conocida de dimensión $n \times J'_k$, $\mathbf{Z}_{jk}$ es una matriz de diseño fija conocida de $n \times q_{jk}$ y $\boldsymbol{\gamma}_{jk}$ es una variable

aleatoria q_{jk}-dimensional dimensional con distribución $N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^{-1})$. $\mathbf{G}_{jk}^{-1}$ es la matriz inversa (generalizada) de una matriz simétrica de $q_{jk} \times q_{jk}$ $\mathbf{G}_{jk} = \mathbf{G_{jk}}(\lambda_{jk})$, la cual puede depender de un vector de hiperparámetros λ_{jk}. Si $\mathbf{G}_{jk}$ es singular, entonces se entiende que γ_{jk} tiene una función de densidad impropia a priori, proporcional a $exp(-\frac{1}{2}\gamma_{jk}^{\mathsf{T}} G_{jk}\gamma_{jk})$, mientras que si no es singular, entonces γ_{jk} tiene una distribución normal q_{jk}-variada con media 0 y matriz de varianza-covarianza $\mathbf{G_{jk}^{-1}}$.

El usuario modelar cada uno de los parámetros de distribución como una función lineal de variables explicativas y/o como funciones lineales de variables estocásticas (efectos aleatorios). Rara vez se tendrán todos los parámetros de distribución para ser modelados utilizando variables explicativas.

Hay varios sub-modelos importantes de los GAMLSS. Por ejemplo, para aquellos que estén familiarizados con el suavizado, la siguiente formulación de un sub-modelo puede ser más familiar. Sea $\mathbf{Z}_{jk} = \mathbf{I}_n$, donde $\mathbf{I}_n$ es la matriz identidad de $n \times n$, y $\gamma_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ para todas las combinaciones de j y k en el modelo, entonces tenemos la formulación *aditiva semi-paramétrica* de GAMLSS dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \textstyle\sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}),$$

donde h_{jk} es una función desconocida de la variable explicativa X_{jk} y $\mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ es el vector el cual evalúa la función h_{jk} en $\mathbf{x}_{jk}$. Si no hubiera término aditivo en ninguno de los parámetros de distribución, tenemos el modelo GAMLSS *lineal paramétrico* simple,

$$g_1(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k,$$

El modelo aditivo semi-paramétrico puede ser extendido para permitir términos paramétricos no-lineales para ser incluidos en el modelo para μ, σ, ν y τ, de la siguiente manera:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) + \textstyle\sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk})$$

Nos vamos a referir al modelo anterior como *aditivo semi-paramétrico no-lineal*. Si, para $k = 1, 2, 3, 4$, $J_k = 0$, esto es, si para todos los parámetros de distribución no tenemos términos aditivos, éste se reduce a un modelo GAMLSS *paramétrico no-lineal*:

$$\mathbf{X}_k^{\mathsf{T}\boldsymbol{\beta}_k} \text{ con } ... = \mathbf{X}_k^{\mathsf{T}}\boldsymbol{\beta}_k$$

Si, adicionalmente, $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) = \mathbf{X}_k^{\mathsf{T}\boldsymbol{\beta}_k}$ para $i = 1, 2, ..., n$ y $k = 1, 2, 3, 4$, entonces el modelo anterior se reduce a un modelo paramétrico lineal. Hay que notar que algunos de los términos en cada $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k)$ pueden ser lineal, en cuyo caso el modelo GAMLSS es

una combinación de términos paramétricos lineales y no-lineales. Vamos a referirnos como modelos GAMLSS paramétricos.

Los vectores paramétricos $\boldsymbol{\beta}_k$ y los parámetros de los efectos aleatorios γ_{jk}, para $j = 1, 2, ..., J_k$ y $k = 1, 2, 3, 4$, son estimados dentro del marco referencial GAMLSS (para valores fijos de los hiperparámetros de suavizado λ_{jk}) mediante la maximización de la función de verosimilitud penalizada $\ell_p(\beta, \gamma)$ dada por

$$\ell_p(\beta, \gamma) = \ell(\beta, \gamma) - \tfrac{1}{2} \sum_{k=1}^{p} \sum_{j=1}^{J_k} \lambda_{jk} \gamma_{jk}^{\mathsf{T}} \mathbf{G}_{jk} \gamma_{jk}$$

donde $\ell(\beta, \gamma) = \sum_{i=1}^{n} log f_Y(y_i|\boldsymbol{\theta}^i) = \sum_{i=1}^{n} log f_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ es la función log-verosimilitud de los parámetros de distribución dados los datos. Notar que se usa (β, γ) como argumento en la log-verosimilitud penalizada para enfatizar que es maximizado; (β, γ) representa todos los $\beta_k's$ y los $\gamma_{jk}'s$, para $j = 1, 2, ..., J_k$ y $k = 1, 2, 3, 4$. Para modelos GAMLSS paramétricos, $\ell_p(\beta, \gamma)$ se reduce a $\ell(\beta)$, y los β_k para $k = 1, 2, 3, 4$ son estimados maximizando la función de verosimilitud $\ell(\beta)$.

2.6. Distribuciones disponibles en GAMLSS

La forma de la distribución asumida por la variable de respuesta Y, $f_Y(y|\mu, \sigma, \nu, \tau)$, puede ser muy general. La única restricción que la implementación en R (R Core Team, 2015) de los GAMLSS tiene es que la función $log f_Y(y|\mu, \sigma, \nu, \tau)$ y su primeras derivadas respecto a cada uno de los parámetros de $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$ deben ser computables. Las derivadas explícitas son preferibles, pero las derivadas numéricas pueden ser usadas.

Los GAMLSS admite la modelización de todos los parámetros de distribución μ, σ, ν y τ como funciones paramétricas lineales o no-lineales y/o funciones de suavizado paramétricas o no-paramétricas de las variables explicativas y/o términos de efectos aleatorios. En la implementación en R, la función *gamlss()* en el paquete `gamlss` (Rigby and Stasinopoulos, 2005) permite fórmulas para todos los parámetros de distribución. Para modelar funciones lineales, se utiliza la formulación de las funciones *lm()*, y *glm()*. Para ajustar funciones no-lineales o no-paramétricas (suavizado) y/o términos de efectos aleatorios, se deben incluir términos aditivos apropiados.

La tabla 2 muestra algunos de los términos aditivos implementados en el paquete `gamlss`.

Distribuciones	No. de parámetros	μ	σ	ν	τ
beta	2	logit	logit	-	-
Box-Cox t	4	identidad	log	identidad	log
exponential Gaussian	3	identidad	log	log	-
generalized beta type 1	4	logit	logit	log	log
skew power exponential type 3	4	identidad	log	log	log
skew t type 2	4	identidad	log	identidad	log
betabinomial	2	logit	log	-	-
Poisson inverse Gaussian	2	log	log	-	-
zero inflated neg. binomial	3	log	log	logit	-
beta inflated (en 0 y 1)	4	logit	logit	log	log

Tabla 1: Tabla de algunas distribuciones disponibles en GAMLSS (con funciones de enlace predeterminadas)

Términos aditivos	nombre de función en R
boosting	boost()
cubic splines based	cs(), scs(), vc()
fractional and power polynomials	fp(), pp()
penalized Beta splines based	pb(), ps(), cy(), tp(), pvc()
random effects	random(), ra(), rc(), re()

Tabla 2: Tabla de algunos términos aditivos implementados en GAMLSS

3. Aplicación

Tal como dijimos en el resumen, al ser analizada una muestra incompleta que no permite determinar una tasa de no respuesta y eventuales sesgos de selección , en este documento se presentan resultados preliminares. Los criterios de selección de los niños fueron los siguientes:

- Niños con examen físico normal al momento del estudio.
- No haber presentado antecedentes luego del 1er año de vida de: sibilancias, asma, broncoespasmo inducido por el ejercicio, y/o bronquitis reiteradas.
- Haber realizado la maniobra de espiración forzada en forma satisfactoria.

Los equipos médicos estaban constituidos por neumólogos Pediatras que realizaban los estudios mediante 2 espirómetros (Brentwood-Spiroscan 2000 y Fukuda) los cuales cumplían con las normas de ATS para estos registros, y enfermeras universitarias. La maestra de la clase del niño estaba presente en los estudios. Se utilizaron piezas bucales descartables para cada niño.

Previamente se pesaron con ropas livianas en una balanza electrónica marca Sohenle Personal Scale 7306.00 (error$\pm$ 0.1kg) y se midió su talla descalzos mediante un pediómetro digital Sohenle 5001 (error $\pm$ 0.5 cm) en un ambiente térmicamente adecuado.

Previamente se habían recabados datos sobre los antecedentes de los niños mediante un formulario escrito enviado a los padres.

3.1. Aspectos Éticos

Se requirió la firma de cada padre aprobando la realización del estudio. Un comité de notables de cada escuela privada aprobó el desarrollo del estudio, explicando previamente el protocolo a seguir. El Consejo de Primaria aprobó la realización del estudio en las escuelas públicas.

4. Resultados

4.1. Descripción de los datos

El tratamiento estadístico de los datos que se detallan a continuación permitirá establecer las funciones respiratorias para ambos sexos y edades.

De un total de 1267 niños participantes, 804 cumplieron con los criterios de inclusión (381 varones y 423 niñas). Sin embrago debe recordarse que los resultados obtenidos en este trabajo serán basados en el anĺisis de aproximandamente 450 niños

Como primera observación (ver figura 3), se puede decir que la relación entre la Talla y el Peso no es necesariamente lineal. También se puede observar que no hay simetría si se toma en cuenta la media local representada por la línea azul, que representa la media local, es decir, no hay homocedasticidad en la variación.

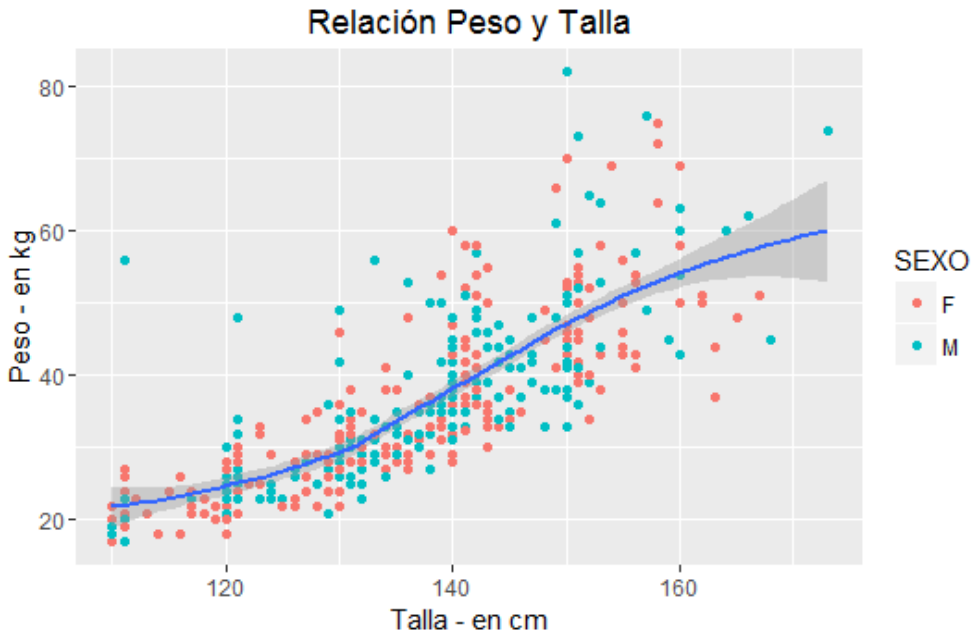

Figura 3: Relación entre la talla y el peso

En el caso del Peso y Edad (figura 4) aparenta ser más lineal. Se observa que respecto a la media local hay una mayor dispersión cuando se incrementa la edad. La heterocedasticidad es mayor.

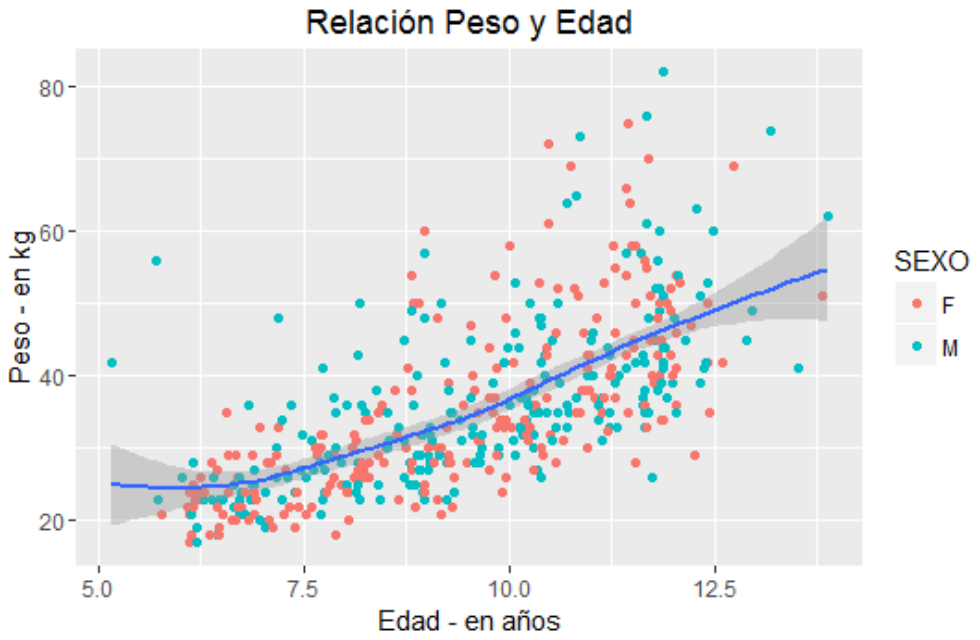

Figura 4: Peso Edad

La relación entre Talla y Edad, (figura 5) exhibe una marcada linealidad, con una menor dispersión que en los gráficos anteriores, aunque se observa cierta heterocedasticidad respecto a la media local.

A continuación se presentan los gráficos que relacionan la variable CVF con las variables explicativas Edad (figura 6), Peso (figura 7) y Talla (figura 8).

Los gráficos (figura 6) y (figura 8) son similares, pareciendo tener una mayor dispersión la variable Talla. No se genera una curva leve en ninguno de los dos casos. Para la variable Peso se observa una concavidad negativa, con una mayor concentración en la primera mitad del gráfico (figura 7).

4.2. Distribución de CVF

El objetivo es encontrar una familia de distribución para la variable de respuesta CVF. Es necesario entonces encontrar una distribución paramétrica que se ajuste a los datos.

Una de las formas para llevar a cabo este cometido es a través de la función *fitdistr()* de la librería `MASS`. El paquete `gamlss`, dentro de sus funciones que trae, incluye la función

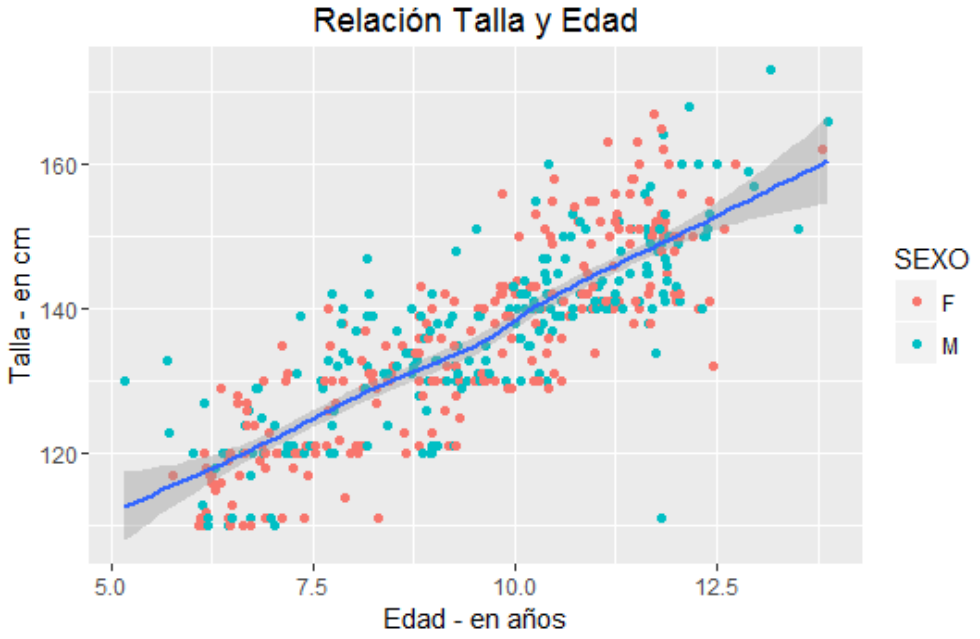

Figura 5: Talla Edad

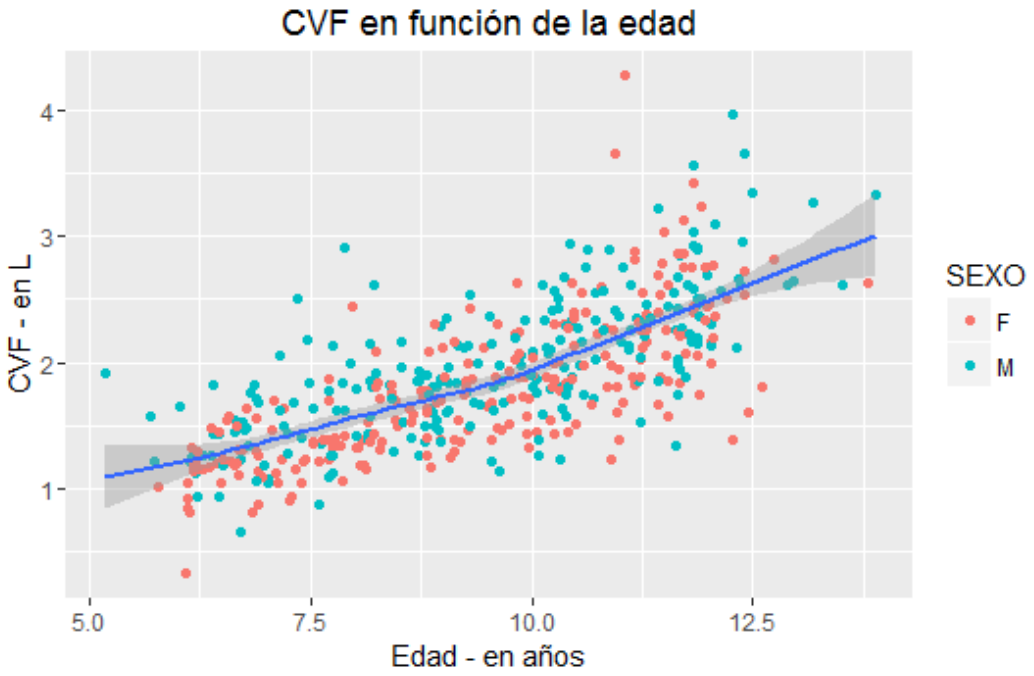

Figura 6: CVF en relación a la Edad

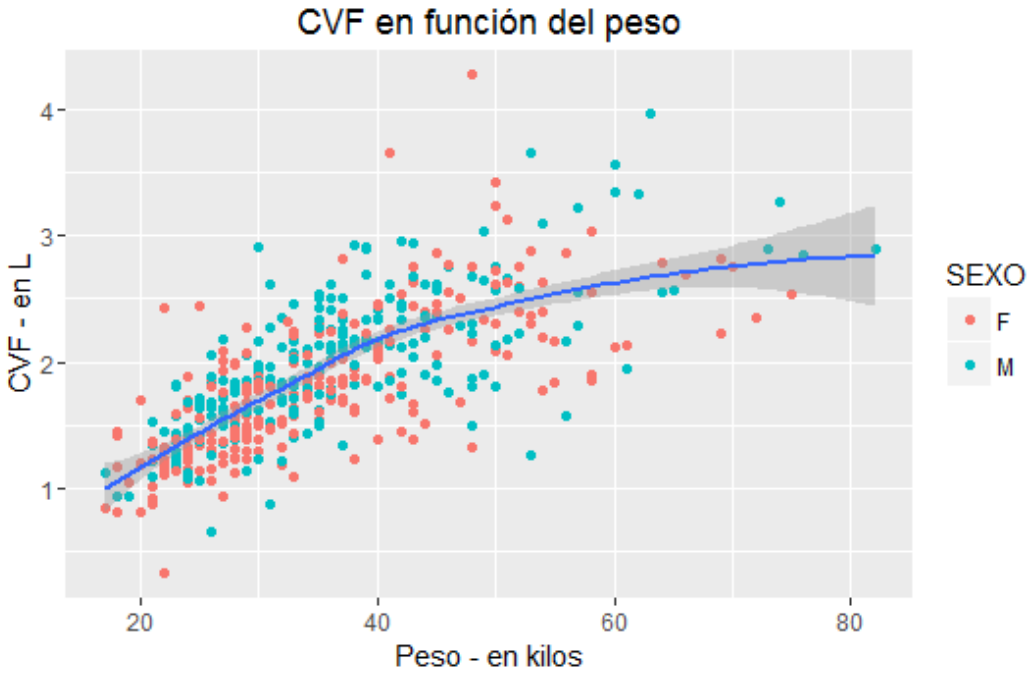

Figura 7: CVF en relación al Peso

fitDist() que sirve como alternativa para el ajuste de una familia de distribución a los datos.

4.2.1. Diferencias entre las funciones *fitDist()* y *fitdistr()*

La gran diferencia entre ambas funciones, es que *fitdistr()*de la librería (`MASS`) (Pinheiro and Bates, 2000) necesita de antemano establecer la distribución (por ejemplo, decirle que es normal) y ésta estima los parámetros de la distribución (media y desvío estándar, para el ejemplo). La función de la librería GAMLSS es mas flexible, ya que a partir de los datos, estima la familia y los parámetros correspondientes.

4.2.2. Función *fitDist* (librería GAMLSS)

La función usa *gamlssML()* permite ajustar todas las funciones de distribución paramétricas relevantes de la gamlss.family a un vector de datos. El modelo final es uno que es seleccionado por el criterio GAIC (Generalized Akaike Information Criterion) con penalización k, que se obtiene agregando una penalización fija a la desvianza global (Global Deviance) (k=2 corresponde al Akaike's Information Criterion, k=log(n) al Schwarz Bayesian

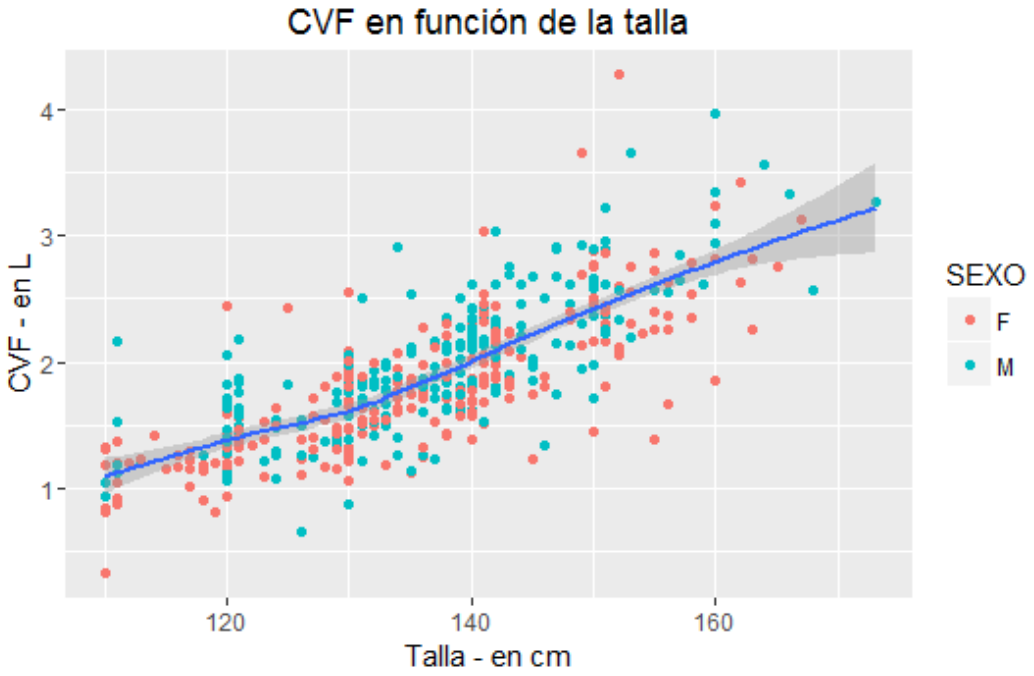

Figura 8: CVF en relación a la Talla

Criterion).

`fitDist(y, k=2, type="realAll", try.gamlss=FALSE, extra=NULL, data=NULL, . . .)`

para incorporar argumentos extra para *gamlssML()* o *gamlss()*.

try.gamlss indica que si el algoritmo no converge, o tiene algún problema con la función *gamlssML()*, que utilice la función *gamlss()* para el ajuste.

```
gamlssML(y, family = NO, weights = NULL, mu.start = NULL,
sigma.start = NULL, nu.start = NULL,tau.start = NULL,
mu.fix = FALSE, sigma.fix = FALSE, nu.fix = FALSE,
tau.fix = FALSE, data = NULL, start.from = NULL, ...)
```

Esta es una función para ajustar una familia de distribución *gamlss* a un conjunto de datos usando un algoritmo de maximización no lineal en R. Esto es relevante solo cuando no hay variables explicativas. De hecho, utiliza la función interna *MLE()* que es una copia de la función *mle()* del paquete `stat4`. La función *gamlssML()* puede ser más rápida para datos grandes que la función equivalente *gamlss()* que es diseñada para modelos de regresión.

4.2.3. Función *fitdistr* (librería `MASS`)

Realiza el ajuste por máxima verosimilitud de distribuciones univariadas, y permite mantener fijos los parámetros que se deseen.

```
fitdistr(x, densfun, start, . . .)
```

parámetros adicionales, tanto para densfun o para optim.En particular, puede ser usado para especificar bordes via lower o upper o ambos.

`dnsfun`: las distribuciones que admiten son las siguientes: beta, cauchy, chi-cuadrado, exponencial, F, gamma, geométrica, log-normal, logística, binomial negativa, normal, Poisson, t y Weibull.
`start`: una lista con los nombres que le asignan a los parámetros a ser optimizados con sus valores iniciales. Para algunas funciones dicha especificación puede ser omitido, pero para otras debe aparecer.

Para las funciones Normal, log-Normal, geométrica, exponencial y Poisson, se usa la forma cerrada de MLE, y no debería usarse la lista con los parámetros.

Para el resto de las distribuciones, se usa la función *optim()* para optimizar la función de log-verosimilitud. Los errores estándar estimados son tomados de la matriz de información observada, calculada por aproximación numérica. Para problemas unidimensionales, se usa el método de Nelder-Mead, y para multidimensionales el método BFGS, a no ser que los argumentos lower o upper sean suministrados (cuando L-BFGS-B es usado) o method es explicitado.

Devuelve la estimación de los parámetros, los errores estándar estimados, la matriz estimada de varianza y covarianza, y el valor de la log-verosimilitud.

4.2.4. Prueba de robustez

Para estudiar la robustez de las distribuciones ajustadas con la función *fitDist()* del paquete `gamlss`, y la dependencia de los resultados en función de los datos, se realiza un proceso de remuestreo iterativo donde se seleccionan muestras de los datos de un tamaño de 400 (cerca del 80 % del total) y se analiza cual es la familia de distribuciones que más se repite en las N iteraciones, y se calcula la variabilidad de los parámetros estimados para cada caso. Se presentan medidas de ajuste basadas en el AIC.

```
iter=1000
nS=400
Results=data.frame(Familia=iter,AIC=0,mu=0,sigma=0,nu=0,tau=0)
for (i in 1:iter) {
sample1=sample_n(na.omit(SubDatosFiltrados), nS)
AjDist=fitDist(CVF, type = "realline", try.gamlss = TRUE, data=sample1)
if (AjDist$df.fit == 4) {
Results[i,]=c(AjDist$family[1],round(AjDist$aic,2),
AjDist$mu, AjDist$sigma,AjDist$nu, AjDist$tau)
}else if (AjDist$df.fit == 3){
Results[i,]=c(AjDist$family[1],round(AjDist$aic,2),
AjDist$mu, AjDist$sigma,AjDist$nu, NA)
}else if (AjDist$df.fit == 2){
Results[i,]=c(AjDist$family[1],round(AjDist$aic,2),
AjDist$mu, AjDist$sigma,NA, NA)}
};
```

Explicación de la rutina para la prueba de robustez.

1. Establece un número de iteraciones en primera instancia.
2. Luego fija un valor de tamaño de muestra del total de datos.
3. Crea un data frame para guardar los resultados de la rutina, donde se va a guardar la familia resultante de la función *fitDist()* (con el menor GAIC, con penalización k=2), luego el valor del AIC, y el valor estimado de los parámetros de la familia de distribución, que van de 2 a 4.
4. Para cada iteración, genera una muestra de tamaño n
5. Aplica la función *fitDist()* a la muestra anteriror para la variable CVF
6. El valor de df.fit, es el equivalente a los grados de libertad del ajuste, que no es otra cosa que la cantidad de parámetros que tiene la familia
7. se almacenan los valores de los cuatro posibles parámetros que pueden caracterizar a la distribución, la familia y el AIC.

En la figura 9 se muestra la frecuencia absoluta de las diferentes familias de distribución propuestas para el ajuste de la variable CVF (sin diferenciar por sexo) con la función *fitDist()*, teniendo en cuenta que para dicha elección el algoritmo de ajuste toma aquella que registra menor GAIC. De un total de 1000 iteraciones, en la tabla 3 se muestra cuantos parámetros tiene dicha familia, la frecuencia absoluta de ajuste, el AIC medio (la media

del AIC del total de repeticiones), y las medias de los parámetros que caracterizan a la distribución.

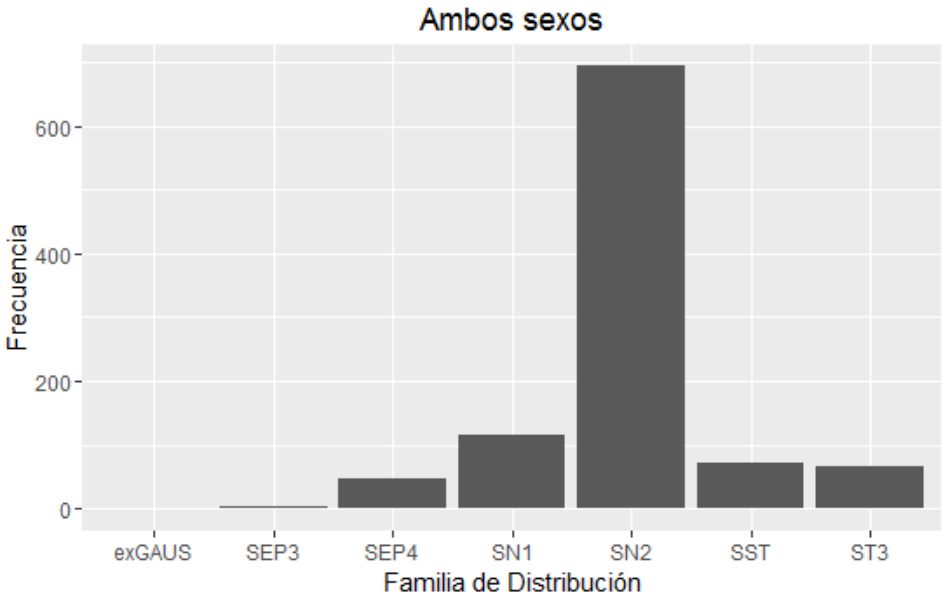

Figura 9: Gráfico de frecuencia de familias de distribución

Familia	n^o parámetros	Frec. Abs	Iteraciones	AIC medio	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\nu}$	$\bar{\tau}$
exGAUS	3	1	1000	661.08	1.50	0.40	0.40	-
SEP3	4	2	1000	654.955	1.44	0.53	1.76	2.36
SEP4	4	48	1000	658.78	1.83	0.79	4.08	1.55
SN1	3	116	1000	675.91	1.24	0.87	2.76	-
SN2	3	694	1000	674.68	1.49	0.50	1.60	-
SST	4	72	1000	672.96	1.88	0.58	1.64	18.11
ST3	4	67	1000	673.69	1.48	0.46	1.64	17.85

Tabla 3: Cuadro de iteraciones para CVF

Para ver gráficamente el ajuste realizado, mediante la función *histDist()*, se grafica la densidad de los datos, la densidad ajustada por la función *fitDist()*, y el histograma, para cada una de las familias de distribución ajustadas, como muestra la figura 10. Para ver con más detalle, aparte se elige graficar la distribución SN2 (Skewed Normal type 2), que fue la que más se repitió, como lo muestra la figura 11.

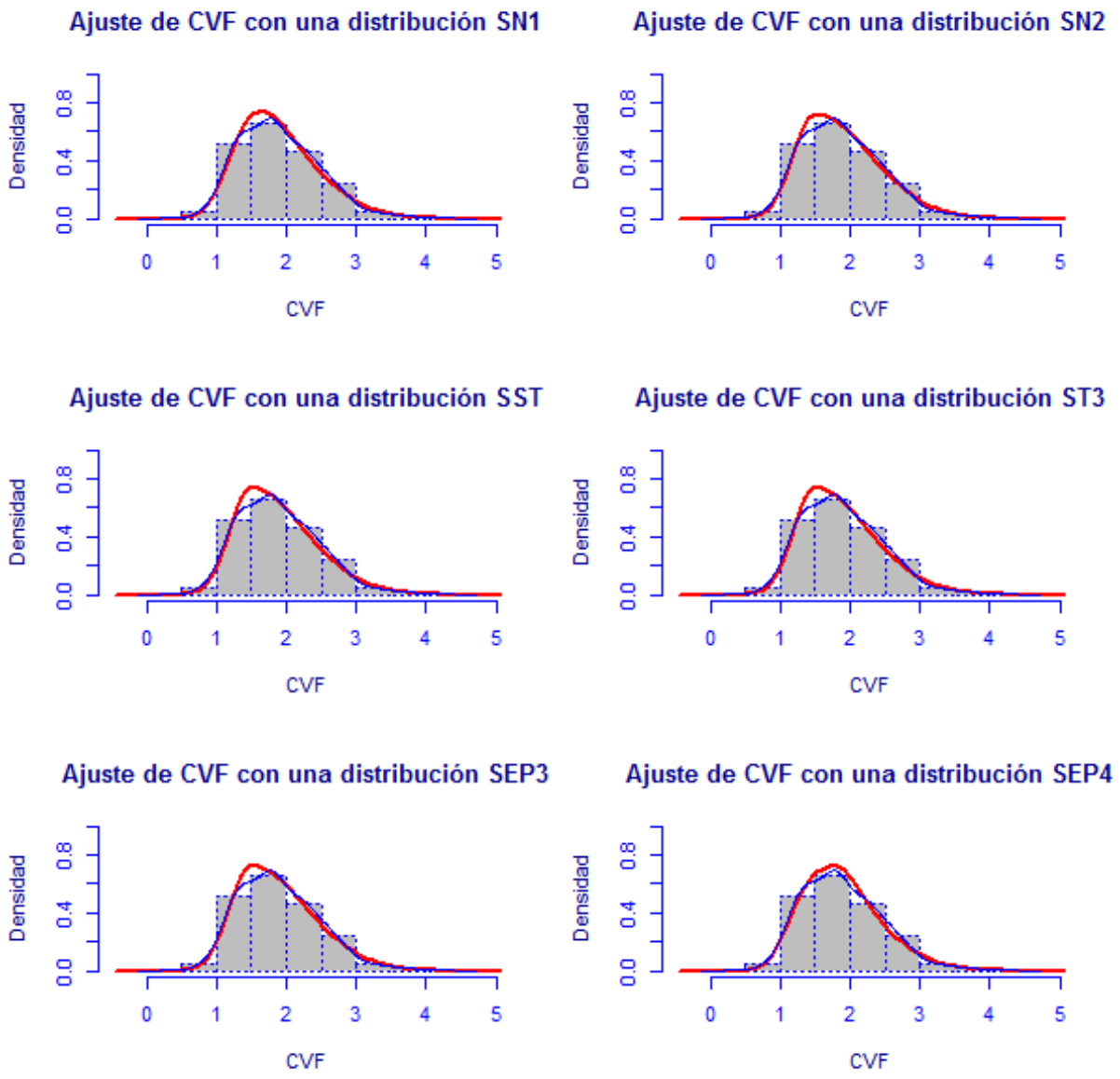

Figura 10: Gráfico de ajustes de densidades

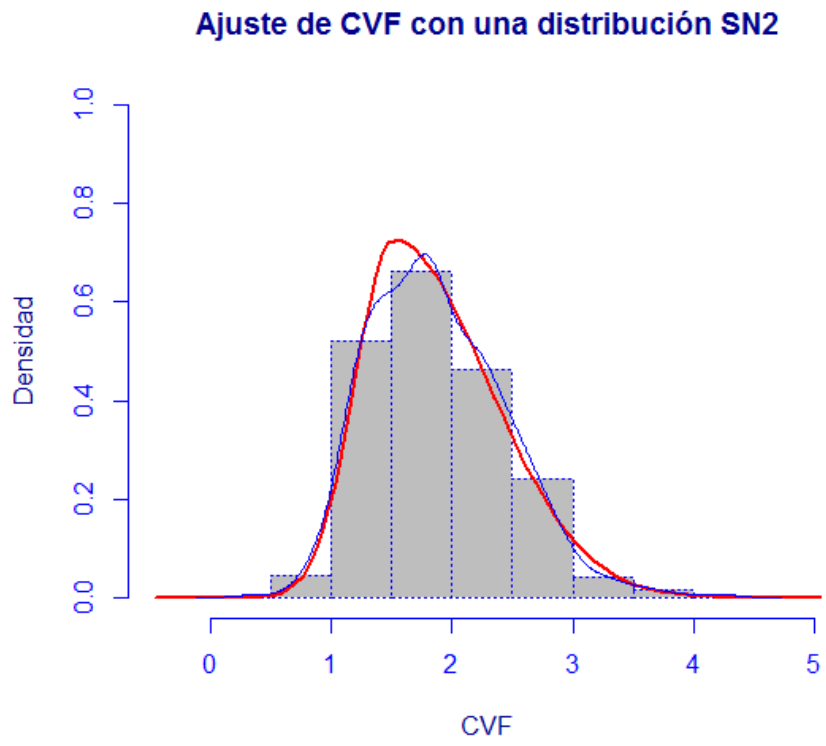

Figura 11: Histograma, densidad empírica y densidad de la familia de distribución SN2

Como se puede apreciar, los ajustes son similares, presentando, a simple vista, leves diferencias entre ellos. De aquí la precaución que se debe tener al utilizar este método de ajuste de distribuciones.

La figura 12 muestra la densidad empírica de la variable de respuesta CVF, en área de color gris claro, y en color rojo la densidad de la distribución log-normal, cuyos parámetros fueron estimados con la función *fitdistr()* del paquete `MASS`.

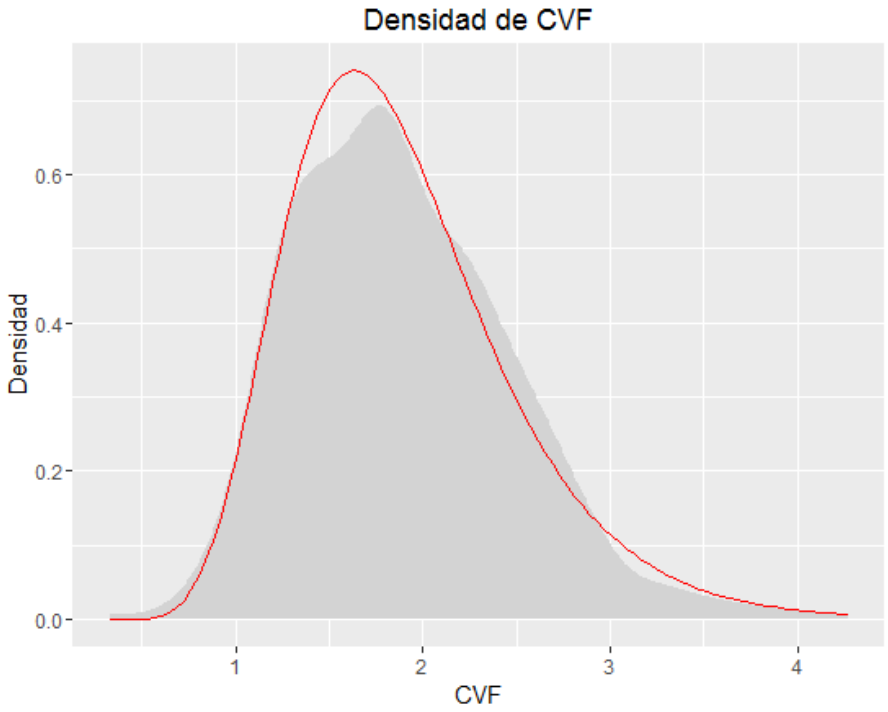

Figura 12: Densidad empírica de los datos y densidad de la distribución log-normal de parámetros estimados por la función *fitdistr()* de la librería `MASS`

5. Discusión

UNa vez analizados los datos preliminares disponibles para este trabajo, se puede decir que , en general , la distribución que mejor se ajusta a la variable CVF ,Teniendo en cuenta la prueba de robustez, es la distribución normal asimétrica (SN2),de tipo II, (skew normal type 2) .

Es importante tener en cuenta que en los resultados considerados no se separan niños y niñas, aspecto importante en los resultados ya que pueden diferir notablemente, porque en realidad al no discriminar se está trabajando con una mezcla de distribuciones (algunos resultados preliminares que aún no se incluyen al dividir por sexo así lo muestran).

Los hallazgos hasta el momento permiten considerar al paquete `gamlss` con una mejor performance que el obtenido usando la librería `MASS` y sus funciones como una alternativa válida a la hora de ajustar datos con densidades "raras" a una distribución paramétrica convencional que puede ser utilizada en modelos.

6. Consideraciones a futuro

- Completar la recolección de los datos y hacer un estudio sobre la No Respuesta y de los eventuales sesgos de selección y evaluar si es necesario hacer algún procedimiento de calibración.
- Plantear algún tipo de distancia que permita comparar y decidir la distribución 'candidata' mas adecuada.
- Estudiar el cambio en los modelos GAM y GAMLSS que se estimen al variar la familia de distribución para la variable de respuesta.
- Estudiar la sensibilidad de los modelos GAM al incorporar términos de mayor orden y modular los factores de penalización que permiten los modelos GAM.
- Comparar los modelos GAM (que no permiten variables del tipo factor) con los GAMLSS dividiendo por sexo y los modelos GAMLSS globales (sin dividir por sexo) que si permite usar variables de tipo factor como regresoras.

Bibliografía

Hastie, T. y Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

Nelder, J. y Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370–384.

Pinheiro, J. y Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rigby, B. y Stasinopoulos, M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

Rigby, B. y Stasinopoulos, M. (2006). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, (6):209–229.

Rigby, B. y Stasinopoulos, M. (2010). A flexible regression approach using GAMLSS in R.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Spriggs, E. (1978). The history of spirometry. *Br.J.Dis.Chest*, 72(3):165–80.

Wood, S. (2006). *Generalized Additive Models : An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL.

Instituto de Estadística

Documentos de Trabajo

Eduardo Acevedo 1139. CP 11200 Montevideo, Uruguay
Teléfonos y fax: (598) 2410 2564 - 2418 7381
Correo: ddt@iesta.edu.uy
www.iesta.edu.uy
Área Publicaciones

Noviembre, 2016
DT-2016/3